Streaming constrained binary logistic regression with online standardized data

Benoît Lalloué, Jean-Marie Monnez, Eliane Albuisson

▶ **To cite this version:**

Benoît Lalloué, Jean-Marie Monnez, Eliane Albuisson. Streaming constrained binary logistic regression with online standardized data. SFC 2019 - XXVIe Rencontres de la Société Francophone de Classification, Sep 2019, Nancy, France. hal-02278090

HAL Id: hal-02278090
https://hal.science/hal-02278090

Submitted on 4 Sep 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Streaming constrained binary logistic regression with online standardized data

Benoît Lalloué[*,**], Jean-Marie Monnez[*,**], Eliane Albuisson[***,****,♯]

[*] Université de Lorraine, CNRS, Inria [1], IECL [2], Nancy, France
[**] CHRU Nancy, INSERM, Université de Lorraine, CIC [3], Plurithématique, Nancy, France
[***] Université de Lorraine, CNRS, IECL [4], Nancy, France
[****] BIOBASE, Pôle S2R, CHRU de Nancy, Vandoeuvre-lès-Nancy, France
[♯] Faculté de Médecine, InSciDenS, Vandoeuvre-lès-Nancy, France
benoit.lalloue@univ-lorraine.fr
jean-marie.monnez@univ-lorraine.fr
eliane.albuisson@univ-lorraine.fr

Abstract. We study a stochastic gradient algorithm for performing online a constrained binary logistic regression in the case of streaming or massive data. Assuming that observed data are realizations of a random vector, these data are standardized online in particular to avoid a numerical explosion or when a shrinkage method such as LASSO is used. We prove the almost sure convergence of a variable step-size constrained stochastic gradient process with averaging when a varying number of new data is introduced at each step. Several stochastic approximation processes with raw data or online standardized data are compared on observed or simulated datasets. The best results are obtained by processes with online standardized data.

1 Introduction

One type of method to analyse streaming or massive data is online learning which proceeds in successive steps, the results of the analysis being updated at each step taking into account a batch of new data. Recursive stochastic algorithms can be used for observations arriving sequentially to estimate for example parameters of a linear regression model (Duarte et al., 2018) or principal components of a factorial analysis (Monnez and Skiredj, 2018) or centres of classes in non-hierarchical clustering (Cardot et al., 2012), whose estimations are updated by each new arriving data batch. In this context, it is not necessary to store the data and, due to the relative simplicity of the computation involved, much more data than with classic methods can be taken into account during the same duration of time. For massive datasets, recursive algorithms can be used by randomly drawing at each step a data batch from the dataset.

Why use online standardized data (each continuous variable is standardized with respect to the estimations at the current step of its expectation and of its standard deviation computed online) and a constrained process? First *to avoid a numerical explosion* as it is studied in Duarte et al. (2018) in the case of sequential multidimensional linear regression. The experiments conducted have shown better performance of processes with online standardized data compared to

those with raw data. Second, *when using a shrinkage method such as LASSO or ridge*, we have first to standardize the explanatory variables. In the case of a data stream, when the mathematical expectation and the variance of each variable are a priori unknown, these variables can be standardized online and a process of the same type can be used but with a projection at each step on the convex set defined by the constraint on the parameters of the regression function. More generally *this type of process can be used for any convex set,* for example if it is imposed that the parameters associated to the explanatory variables are positive. Third we can consider the case where a logistic model with standardized explanatory variables is defined and *where explanatory variables have an expectation and a variance that may depend on time or on the values of controlled variables* according to a specific model; this assumes that we can estimate online the expectation and the variance of these variables.

A suitable choice of step-size is often crucial for obtaining good performance of a stochastic gradient process. If the step-size is too small, the convergence will be slower. Conversely, if it is too large, a numerical explosion may occur during the first iterations. We use here *an averaged stochastic gradient process, with a piecewise constant step-size* as suggested in Bach (2014) in order that the step-size does not decrease too quickly and reduces the speed of convergence.

2 Study of a stochastic gradient process

Suppose that data are realizations of a random vector $(R^1, ..., R^p, S)$ in $\mathbb{R}^p \times \{0, 1\}$.

Let A' be the transpose of a matrix A. Let R be the random column vector $\left(R^1 \ ... \ R^p \ 1\right)'$, $m = \left(E\left[R^1\right] \ ... \ E\left[R^p\right] \ 0\right)'$, $R^c = R - m$, r^c a realization of R^c, σ^k the standard deviation of R^k, $k = 1, ..., p$, Γ the diagonal $(p+1, p+1)$ matrix with diagonal elements $\frac{1}{\sigma^1}, ..., \frac{1}{\sigma^p}, 1$ (taking by convention $\sigma^k = 1$ for a discrete variable), $Z = \Gamma R^c$, whose continuous components are standardized, $z = \Gamma r^c$ a realization of Z, $\theta = \left(\theta^1 \ ... \ \theta^P \theta^{P+1}\right)'$ a column vector of real parameters.

Consider the logistic model with standardized covariates:

$$P\left(S = s \mid R = r\right) = f(s; z, \theta) = \left(\frac{e^{z'\theta}}{1 + e^{z'\theta}}\right)^s \left(\frac{1}{1 + e^{z'\theta}}\right)^{1-s} = \frac{e^{z'\theta s}}{1 + e^{z'\theta}}.$$

$E\left[S \mid R\right] = h\left(Z'\theta\right)$ with $h(u) = \frac{e^u}{1+e^u} = \frac{1}{1+e^{-u}}$.

Define the loss function -ln $f\left(s; z, x\right) = -z'xs + \ln\left(1 + e^{z'x}\right)$. The cost function

$$F(x) = -E\left[\ln f\left(S; Z, x\right)\right] = E\left[-Z'xS + \ln\left(1 + e^{Z'x}\right)\right]$$

has θ for unique minimizer since F is a convex function with positive hessian. θ is the unique solution of:

$$F'(x) = E\left[-ZS + \frac{Ze^{Z'x}}{1 + e^{Z'x}}\right] = E\left[Z\left(h\left(Z'x\right) - S\right)\right] = 0.$$

Let $\left((R_n^1, ..., R_n^p, S_n), n \geqslant 1\right)$ be an i.i.d. sample of $(R^1, ..., R^p, S)$, for $n \geqslant 1$, $R_n = \left(R_n^1 \ ... \ R_n^p \ 1\right)'$, for $k = 1, ..., p$, $\overline{R}_n^k$ the mean of the sample $\left(R_1^k, ..., R_n^k\right)$ of R^k and $\left(V_n^k\right)^2 =$

$\frac{1}{n}\sum_{i=1}^{n}\left(R_i^k - \overline{R_n^k}\right)^2$ its variance, both recursively computed, $\overline{R}_n = \left(\overline{R_n^1}\ ...\ \overline{R_n^p}\ 0\right)'$ and Γ_n the diagonal $(p+1, p+1)$ matrix with diagonal elements $\frac{1}{\sqrt{\frac{n}{n-1}V_n^1}}, ..., \frac{1}{\sqrt{\frac{n}{n-1}V_n^p}}, 1$.

Suppose that m_n observations (R_i, S_i) are taken into account at step n of the defined process. Let $\mu_n = \sum_{i=1}^{n} m_i, I_n = \{\mu_{n-1}+1, ..., \mu_n\}$, $\widehat{R}_n = \overline{R}_{\mu_n}$, $\widehat{\Gamma}_n = \Gamma_{\mu_n}$ and for $j \in I_n, \widetilde{Z}_j = \widehat{\Gamma}_{n-1}\left(R_j - \widehat{R}_{n-1}\right)$: for $k = 1, ..., p$, each component R_j^k of R_j is pseudo-standardized with respect to the empirical mean $\widehat{R}_{n-1}^k$ and to the empirical estimation of σ^k, $\sqrt{\frac{n}{n-1}V_{\mu_{n-1}}^k}$.

Suppose that θ is constrained to belong to a convex subset K of $\mathbb{R}^{p+1}$. Let Π be the projection operator on K. Recursively define the stochastic approximation processes (X_n) of the Robbins-Monro type (Robbins and Monro, 1951) and $\left(\overline{X}_n\right)$ in $\mathbb{R}^{p+1}$:

$$X_{n+1} = \Pi\left(X_n - a_n \frac{1}{m_n}\sum_{j\in I_n}\widetilde{Z}_j\left(h\left(\widetilde{Z}_j' X_n\right) - S_j\right)\right),\ \overline{X}_{n+1} = \frac{1}{n+1}\sum_{1}^{n+1} X_i.$$

Theorem 1 *Suppose there is no affine relation between the components of R, the moments of order 4 of R exist and* $a_n > 0$, $\sum_{n=1}^{\infty} a_n = \infty$, $\sum_{n=1}^{\infty} \frac{a_n}{\sqrt{n}} < \infty$, $\sum_{n=1}^{\infty} a_n^2 < \infty$. *Then* (X_n) *and* $\left(\overline{X}_n\right)$ *converge to* θ *a.s.*

The proof is in Lalloué et al. (2019b).

3 Experiments

Stochastic approximation processes were compared, including classic stochastic gradient descent (SGD) with a variable step-size, averaged stochastic gradient descent (ASGD) with a piecewise constant step-size with level sizes 50, 100 or 200, and the same processes but with online standardization of the data (Section 2). For these 8 processes, 3 variants with 1, 10 or 100 new observations per step were tested. Therefore 24 processes are studied. For processes with a variable step-size, we have defined $a_n = \frac{c}{(b+n)^\alpha}$, for those with a piecewise constant step-size, $a_n = \frac{c}{(b+\lfloor\frac{n}{\tau}\rfloor)^\alpha}$ where $\lfloor . \rfloor$ denotes the integer part and τ is the size of the levels. We set $\alpha = 2/3$, $b = 1$, $c = 1$. All processes were initialized with $X_1 = 0$.

We used as "gold standard" the vector of coefficients θ^c obtained by classic logistic regression (using R's glm function). Let $\hat{\theta}_{n+1}$ be the estimated vector obtained by a tested process after n iterations. The cosine of the angle between θ^c and $\hat{\theta}_{n+1}$ was used as a convergence criterion: $cos(\theta^c, \hat{\theta}_{n+1}) = \frac{\theta^{c\prime}\hat{\theta}_{n+1}}{\|\theta^c\|\|\hat{\theta}_{n+1}\|}$.

The processes were tested on five datasets available on the Internet (Twonorm, Ringnorm, Quantum, Adult, EEG) and the HOSPHF30D dataset derived from the EPHESUS study (Pitt et al., 2003), all already used to test the performance of processes with online standardized data in the case of online linear regression (Duarte et al., 2018). Twonorm and Ringnorm contain

simulated data. Adult, EEG and HOSPHF30D contain observed data with outliers, variables of different types and scales, unlike Quantum.

At each step of a process a data batch is randomly drawn from the dataset. All processes were applied on all datasets for a fixed number of observations used and for a fixed processing time (the cumulative time to compute the process updates, excluding operations such as data sampling, data management, formatting and recording of results). As an example, for a processing time of 60s (Figure 1) all tested processes using raw observed data, except Quantum, had a numerical explosion. Abbreviations used in Figure 1 are: C for classic SGD or A for ASGD, R for raw data or S for online standardized data, V for variable step-size or P for piecewise constant step-size; for instance, AR1P50 is the averaged process with raw data, 1 observation per step, piecewise constant step-size with level size 50, CS1V is the classic process with online standardized data, 1 observation per step and variable step-size.

Process	Twonorm	Ringnorm	Quantum	Adult	EEG	HOSPHF30D	Mean rank
CR1V	0.9999	0.9999	0.9709	EXPL	EXPL	EXPL	20.8
CR10V	1.0000	1.0000	0.9683	EXPL	EXPL	EXPL	21.8
CR100V	1.0000	1.0000	0.9659	EXPL	EXPL	EXPL	22.5
AR1P50	1.0000	1.0000	0.9978	EXPL	EXPL	EXPL	19.3
AR10P50	1.0000	1.0000	0.9960	EXPL	EXPL	EXPL	18.3
AR100P50	1.0000	1.0000	0.9948	EXPL	EXPL	EXPL	20.0
AR1P100	1.0000	1.0000	0.9991	EXPL	EXPL	EXPL	18.0
AR10P100	1.0000	1.0000	0.9972	EXPL	EXPL	EXPL	17.5
AR100P100	1.0000	1.0000	0.9959	EXPL	EXPL	EXPL	19.0
AR1P200	1.0000	1.0000	0.9999	EXPL	EXPL	EXPL	16.8
AR10P200	1.0000	1.0000	0.9981	EXPL	EXPL	EXPL	16.5
AR100P200	1.0000	1.0000	0.9970	EXPL	EXPL	EXPL	18.2
CS1V	0.9997	0.9998	0.9987	0.9979	0.9988	0.9898	17.5
CS10V	0.9996	1.0000	0.9989	0.9968	0.9994	0.9932	15.8
CS100V	0.9994	1.0000	0.9992	0.9953	0.9993	0.9840	15.3
AS1P50	0.9999	1.0000	0.9959	0.9964	0.9993	0.9854	17.2
AS10P50	1.0000	1.0000	0.9999	0.9998	0.9997	0.9986	8.2
AS100P50	1.0000	1.0000	0.9999	0.9999	1.0000	0.9999	6.5
AS1P100	0.9999	1.0000	0.9948	0.9888	0.9992	0.9841	19.2
AS10P100	1.0000	1.0000	0.9999	0.9998	0.9996	0.9987	8.8
AS100P100	1.0000	1.0000	0.9999	0.9999	1.0000	0.9999	6.7
AS1P200	0.9999	0.9999	0.9934	0.9823	0.9987	0.9812	19.8
AS10P200	1.0000	1.0000	0.9999	0.9996	0.9996	0.9986	8.2
AS100P200	1.0000	1.0000	0.9999	1.0000	1.0000	0.9999	4.8

EXPL: numerical explosion.
Process type: C for classic SGD, A for ASGD. *Data type:* R for raw data, S for online standardized data.
First number: number of new observations at each step.
Step-size: V for variable, P for piecewise constant (*second number* is the levels size).

FIG. 1 – *Cosines for 1 minute of processing time*

For each dataset and at each recording point (see below), processes were ranked from the best (highest cosine) to the worst (lowest cosine). The mean rank over all datasets was used to compare the processes at a given recorded point and globally. Over all datasets, the processes with the best results after 60s are averaged processes with online standardization and

piecewise constant step-sizes, the best one with levels of size 200 and 100 new observations per step (AS100P200).

As in Duarte et al. (2018), the values of the criterion for each process were recorded every N observations used from N to $100 \times N$, N being the number of observations in a dataset, and every second of processing time from 1 to 120s. As an example, when studying the evolution of the rankings with the processing time, two groups of processes appear clearly from the beginning and remain during all the studied period. The group with the worst rankings (at the top in Figure 2) contains all processes using raw data, all processes using only one new observation at each step, and all "classic" processes. The group with the best rankings (at the bottom in Figure 2) contains all averaged processes with online standardization, piecewise constant step-sizes, and using 10 or 100 new observations per step, the best one with levels of size 200 and 100 new observations per step. Other results can be found in Lalloué et al. (2019b).

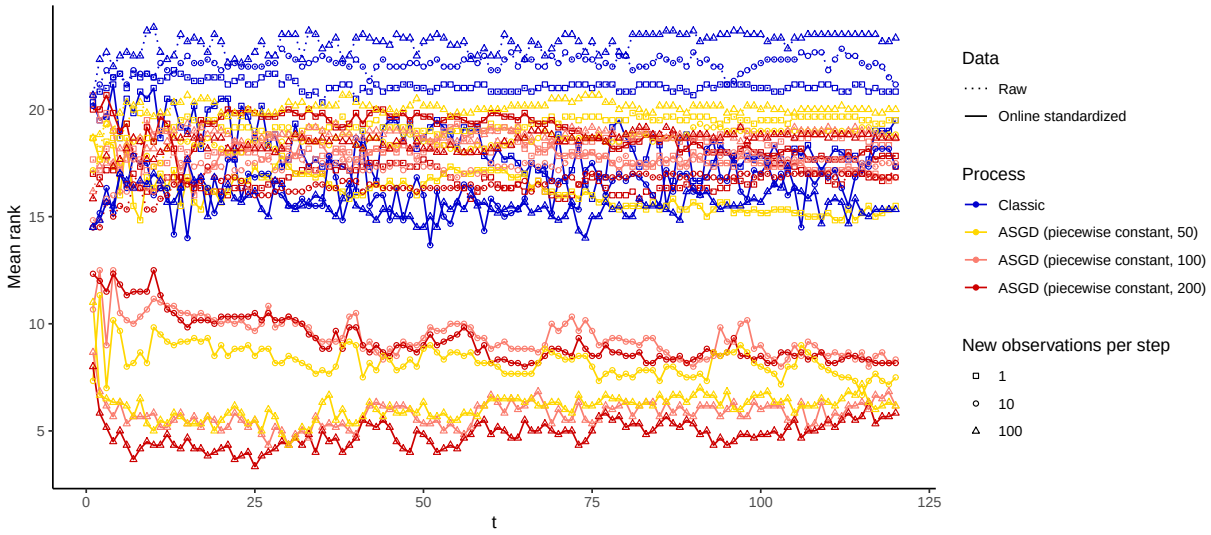

FIG. 2 – *Evolution with the processing time*

Conclusion

We have studied an averaged constrained stochastic gradient algorithm for performing online a constrained binary logistic regression. We have proposed to use an online standardization of the data to avoid a numerical explosion, or when a shrinkage method (such as LASSO) is used, or even when expectations or variances of explanatory variables change (varying with time or depending on the values of controlled variables) and can be estimated online. We have proposed to use a decreasing piecewise constant step-size in order that it does not decrease too quickly and consequently reduces the speed of convergence of the process. We have made experiments on observed and simulated datasets. The results confirm the validity of the choices

made: online standardization of the data, averaged process and piecewise constant step-size. This algorithm is used for scoring online heart failure (Lalloué et al., 2019a).

Acknowledgement

Results incorporated in this article received funding from the investments for the Future Program under grant agreement No ANR-15-RHU-0004.

References

Bach, F. (2014). Adaptivity of averaged stochastic gradient descent to local strong convexity for logistic regression. *Journal of Machine Learning Research 15*, 595–627.

Cardot, H., P. Cénac, and J.-M. Monnez (2012). A fast and recursive algorithm for clustering large datasets with k-medians. *Computational Statistics & Data Analysis 56*(6), 1434–1449.

Duarte, K., J.-M. Monnez, and E. Albuisson (2018). Sequential linear regression with online standardized data. *PLOS ONE 13*(1), e0191186.

Lalloué, B., J.-M. Monnez, and E. Albuisson (2019a). Actualisation en ligne d'un score d'ensemble. In *51e Journées de Statistique*, Nancy, France. Société Française de Statistique. *hal-02152352*.

Lalloué, B., J.-M. Monnez, and E. Albuisson (2019b). Streaming constrained binary logistic regression with online standardized data. Application to scoring heart failure. *hal-02156324*.

Monnez, J.-M. and A. Skiredj (2018). Convergence of a normed eigenvector stochastic approximation process and application to online principal component analysis of a data stream. *hal-01844419*.

Pitt, B., W. Remme, F. Zannad, J. Neaton, F. Martinez, B. Roniker, R. Bittman, S. Hurley, J. Kleiman, and M. Gatlin (2003). Eplerenone, a selective Aldosterone blocker, in patients with left ventricular dysfunction after myocardial infarction. *New England Journal of Medicine 348*(14), 1309–1321.

Robbins, H. and S. Monro (1951). A stochastic approximation method. *The Annals of Mathematical Statistics 22*(3), 400–407.

Résumé

Nous étudions un algorithme de gradient stochastique pour réaliser une régression logistique sous contraintes dans le cas de données massives ou en ligne. En supposant que les données observées sont les réalisations d'un vecteur aléatoire, ces données sont standardisées en ligne pour éviter une explosion numérique ou lorsqu'une méthode de pénalisation telle que LASSO est utilisée. Nous démontrons la convergence presque sûre d'un processus de gradient stochastique moyenné à pas variable lorsqu'un nombre variable de nouvelles données sont introduites à chaque étape. Vingt-quatre processus d'approximation stochastique avec des données brutes ou standardisées en ligne sont comparés sur des données réelles ou simulées. Les meilleurs résultats sont obtenus pour des processus avec données standardisées.